**Fourth: Communicate with Stakeholders**

Subject: Data quality issue report with Receipts, Brands, and Users data

Dear Sir/Madam (replaced by the product/business leader name),

After careful exploration on Receipts, Brands and Users datasets, I propose the following suggestions so that our company could have a more comprehensive and tidier dataset for future analysis to guide our business decision-making:

1. Change the current data schema of Receipts: rewardsReceiptItemList
2. Collect more complete datasets and decrease missing values
3. Reduce data redundancy in Users
4. Correct the anomalous relationship of barcode in Brands and timestamp in Receipts
5. Verify the correctness of abnormal values in Receipts
6. Re-design the category and category code
7. Change the date format to mm/dd/yyyy

And my reasons are as follows:

1. In current Receipts data, the Data format of rewardsReceiptItemList is not consistent, so it is not able to process in batch. For example, some items only have 6 features such as description, quantity, price, etc., while some other items have 18 features. This will lead to a lot of missing values when we want to structure the data and make effective analysis.
2. All the three dataset includes certain amounts of missing values. However, there are a significant number of missing values in Receipts and Brands. With the missing values over certain percentage, these features will not provide enough information. I have listed the attributes of the two tables with missing value percentage here:

Table : Receipts Data

|  |  |  |
| --- | --- | --- |
|  | Number of missing values | Missing value percentage (%) |
| **pointsAwardedDate** | 582 | 52.01 |
| **bonusPointsEarnedReason** | 575 | 51.39 |
| **bonusPointsEarned** | 575 | 51.39 |
| **finishedDate** | 551 | 49.24 |
| **pointsEarned** | 510 | 45.58 |
| **purchasedItemCount** | 484 | 43.25 |
| **purchaseDate** | 448 | 40.04 |
| **rewardsReceiptItemList** | 440 | 39.32 |
| **totalSpent** | 435 | 38.87 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table : Brands Data   |  | Number of missing values | Missing value percentage (%) | | --- | --- | --- | | **categoryCode** | 650 | 55.70 | | **topBrand** | 612 | 52.44 | | **brandCode** | 234 | 20.05 | | **category** | 155 | 13.28 | |  |  |

1. In the total 495 records of Users data, only 42.8% User\_id is unique. Because my team need to merge data with other tables, the redundancy in Users data will lead to duplications in the merged table and thus we should avoid in advance.
3. The barcode in Brands is used as identifier to merge data. However, I found that some products with same barcode represent different products. For example, barcode 511111605058 represents magazine brand and dairy brand at the same time. And it will lead to confusions when joining data
4. After transforming timestamp in Receipts to date format, I found that some dates scanned are 1 month previous to the dates purchased. It is obviously against our business logic: customers made purchase first and then scan receipts into our system. And it should be fixed.
5. According to my box plots, I found that some abnormal values in purchased item count and points earned.
6. if the purchased item count is 0, there should not be a receipt exists and the record is invalid.
7. for the points earned, some customers earned points over 5000 with one receipt. It’s not very common in reality, especially with spending less than $100. It’s potential fraud in this case.
8. one customer scanned 436 times, which is significantly higher than the other customers. It’s better to check this customer’s record in case of fraud.
9. Currently, there are some duplications in brand category. For example, “Dairy”, “ Dairy & Refrigerated” are same categories. In category code, the duplications are eliminated. However, the category code misses some categories, such as ”Canned Goods & Soups” and “Deli”. I recommend that business team design the category and category code again to solve the issues.
10. Current date format is UNIX Timestamp, which is hard to extract information. I suggest changing it to mm/dd/YYYY format or similar format so it can smooth date processing.

Please feel free to let me know if you have any questions. I’m happy to schedule a meeting to discuss these solutions in detail at your convenience.

Sincerely,

Shuwen Huang